Final Project Report

Diabetes Prediction Using Machine Learning

1. Project Overview

This project aims to develop a machine learning model that can predict whether an individual has diabetes based on medical and demographic features. With the growing prevalence of diabetes globally, early prediction can enable timely intervention and reduce long-term health risks.

# 2. Motivation and Problem Statement

Diabetes is a life-threatening condition affecting over 500 million people worldwide. Early detection can help prevent severe complications such as kidney failure and cardiovascular diseases. The goal of this project is to use machine learning (ML) to predict diabetes presence using patient information.

# 3. Dataset Description

The dataset, sourced from Kaggle, contains anonymized clinical and demographic data for diabetes prediction. It includes features such as age, gender, BMI, hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. The target variable is a binary indicator of diabetes presence.

# 4. Project Workflow

The steps undertaken include:

**- Data Collection**

**- Data Cleaning and Preprocessing**

**- Exploratory Data Analysis (EDA)**

**- Feature Selection**

**- Model Building and Training**

**- Model Evaluation and Validation**

# 5. Data Preprocessing

We dropped rows with missing values and encoded categorical variables. Gender was mapped to numerical values (Male: 0, Female: 1, Other: 2) and one-hot encoding was applied to 'smoking history'. MinMaxScaler was used to normalize numerical features.

# 6. Exploratory Data Analysis (EDA)

EDA was conducted to understand data distribution and feature relationships. Key insights:

- Strong correlation found between diabetes and HbA1c, blood glucose levels.

A diagram of a heatmap

AI-generated content may be incorrect.

- Visualizations included histograms, boxplots, and heatmaps.

A graph of a level distribution

AI-generated content may be incorrect.

A diagram of a diabetes status

AI-generated content may be incorrect.

# 7. Feature Selection

Selected features included HbA1c level and blood glucose (strong predictors), as well as BMI and age (moderate predictors). All features were retained for initial model comparison.

# 8. Model Building and Training

Three classification models were trained and tested:

**1. Logistic Regression:** Interpretable and fast. Baseline for performance comparison.

**2. Decision Tree Classifier:** Handles nonlinear relationships. Can be overfit without pruning.

**3. Random Forest Classifier:** Ensemble of decision trees. Best performance among all models. Robust to overfitting and handles feature importance well.

* Random Forest outperformed others across most metrics.

A screenshot of a computer screen

AI-generated content may be incorrect.

# 9. Model Evaluation

Models were evaluated using the following metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Random Forest provided the highest AUC and F1-score, indicating good balance between precision and recall.

# 10. Challenges and Limitations

Limitations include:

- Use of simulated data instead of real clinical records

- Potential class imbalance

- Lack of lifestyle-related features like diet and physical activity

# 11. Conclusion

The machine learning models demonstrated effective prediction capabilities, particularly Random Forest. Visual analysis reinforced known medical associations.

# 12. Future Work

Future enhancements can include:

- Collecting real-world clinical data

- Adding lifestyle features

- Exploring advanced models and deployment as a web app